IBM S/390 Parallel Enterprise Servers G3 and G4

by G. S. Rao
T. A. Gregg
C. A. Price
C. L. Rao
S. J. Repka

This overview paper describes the key steps taken by IBM to transform the S/390® mainframe platform and to enhance customer satisfaction with improvements in cost, scalability, and application enablement. The effectiveness of the transformation is discussed in the context of performance and reliability, and the significance of cluster architecture is defined. Finally, mainframe resurgence is discussed, and factors important to enabling the growth of servers and microprocessors are presented.

Introduction: Redefining the S/390 mainframe

• *System structure requirements for the S/390 platform transition*

In 1993, IBM began revitalizing the S/390* platform in several steps, with most of the transformation complete by mid-1997; key elements of this process included the following:

• *Reducing the cost of the processor by switching to CMOS chip technology* This made possible price/performance improvements of 35–40% per year. Since the performance capability of CMOS would not accommodate customer workloads currently running on bipolar (ES/9000*) systems, the capacity of the symmetric multiprocessor (SMP) system and the engine size of CMOS systems were increased each year to provide replacements for successive generations of bipolar systems (**Figure 1**).

• *Introduction of parallel systems* To meet customer requirements for higher availability, incremental increases in capacity, and large-capacity scaling, IBM introduced a unique clustering approach for S/390 (the Parallel Sysplex*) [1, 2]. This approach confined the need to adapt to the cluster technology to the operating system and the database and transaction monitoring systems. This technology is now supported in all key IBM-supplied operating systems and compilers and many independent software vendor products.

• *Support of open interfaces* This has included support for connection interfaces such as Ethernet, FDDI, and ATM; offering TCP/IP as an alternative network protocol to SNA; and supporting UNIX** standard programming interfaces.

Effectiveness of bipolar replacement

• *Performance*

In the transition of S/390 systems from bipolar to CMOS technology, rapid improvements in the circuit density and

0018-8646/97/$5.00 © 1997 IBM

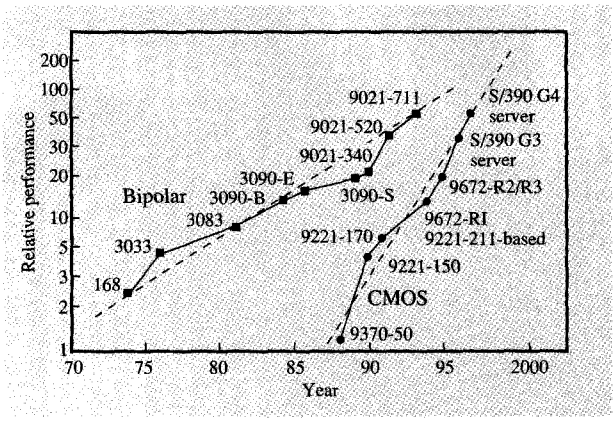


Figure 1

S/390 uniprocessor performance.

performance of CMOS have allowed IBM to produce four generations of S/390 CMOS systems in the past four years (1994–1997). The uniprocessor "size," or effective throughput, of the last three generations (9672-R*2, 9672-R*3, 9672-R*4, and 9672-R*5) delivers uniprocessor speeds roughly comparable to those of the bipolar versions of the 3090-180J, 9021-520, and 9021-711 uniprocessors, respectively, as shown in **Table 1**.

- *Reliability*

In addition to the rapid improvement in processing capacity of the S/390 CMOS-based systems, an even more dramatic improvement has been made in the reliability of the systems. Reliability enhancement through a) improved intrinsic failure rates of CMOS technology versus bipolar; b) extensive reductions in the total number of parts required for the CMOS systems (**Figure 2**); and c) continued improvements in fault-tolerant design [redundancy, ECC, cache-line delete, sparing, and $N + 1$ power (an extra power supply)] resulted in an improvement of nearly two orders of magnitude in MTBF (mean time between failures) of the 9672-RX5 over a 3090-600J bipolar system.

Scalability of performance

Regardless of processor technology, the scalability of a multiple-processor system representing a single-system image (SSI) is limited to a small number of processors. This is due to the well-known fact that as the size of the single system increases, the cost of the hardware and software components of the system increases prohibitively in order to sustain scalability for a large number of processors. The IBM Parallel Sysplex constitutes multiple computing images operating under OS/390* in a system complex (sysplex). The sysplex uses the coupling facility as a means of sharing data, resulting in a high-performance parallel OS/390 operating system configuration. With the advent of this IBM parallel sysplex technology, several processor complexes, each in its own right a multiple processor, could be made to appear as a single system so far as the end user is concerned. With this transition to parallel sysplex technology, it is also possible to maintain a near-linear degree of scalability even with the use of several tens of processors. **Figure 3** shows this desirable scalability characteristic for the parallel sysplex in comparison to the scalability of a single multiple-processor system. In this study, the LSPR (Large Systems Performance Reference) mixed workload[1] is run on the single multiprocessor system, pertaining to either bipolar or CMOS technology.

Table 1 Equivalence of CMOS and bipolar uniprocessors.

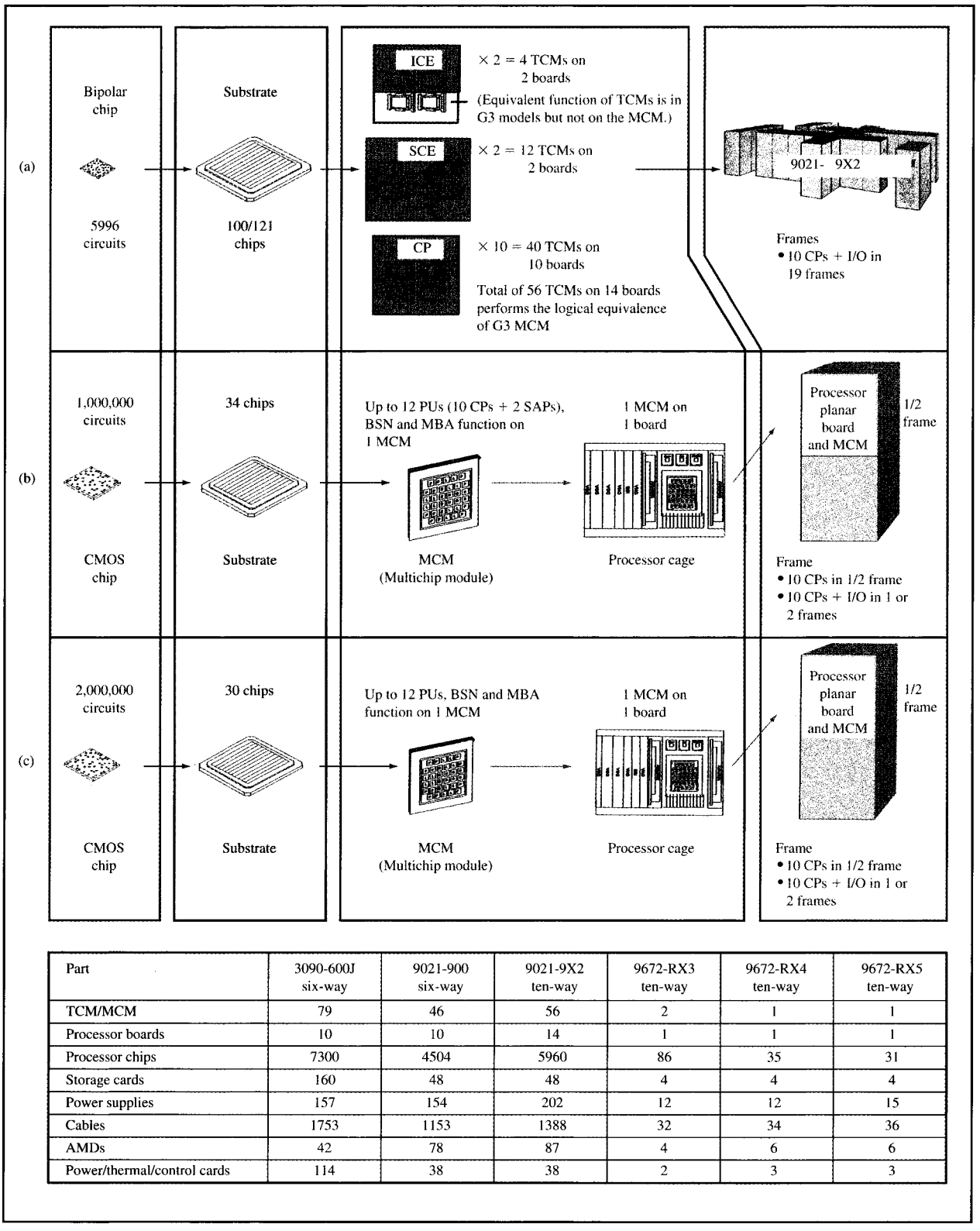
Date	*Generation*	*CMOS model*	*Equivalent bipolar model*
4/94	G1	9672 R11	3090-180
7/95	G2	9672 R12	3090-180J
9/96	G3	9672 R14	9021-520
6/97	G4	9672 R15	9021-711

The parallel sysplex comprising CMOS engines is used to run either CICS/DBCTL or IMS/DB2 data-sharing EBM benchmarks. In the CICS/DBCTL workload, the IBM CICS* subsystem acts as the transaction manager, and the database manager is the IMS DBCTL. This workload contains transactions from CICS applications. In the IMS/DB2 workload, the transaction manager is IMS* at the front end, with DB2* V4 acting as the database manager. The degree of data sharing in both of the workloads is maintained at 100%. The scalability trend is approximately the same for the two workloads, and continues to hold when the parallel sysplex of CMOS G1 is replaced with G2, G3, or G4. There is an initial cost for customers of about 10% in throughput when two systems are enabled for data sharing in comparison to a non-data-sharing single system. As more systems join the parallel sysplex, each additional system incurs less than 0.5% cost in throughput, thereby providing a high degree of scalability. Thus, S/390 parallel sysplex scalability supports a very large commercial data processing environment.

Continuous availability design point

In addition to the high-availability characteristics described above which protect customers from unscheduled outages, many other system design features

[1]The LSPR mixed workload is composed of equal proportions of commercial batch, TSO, CICS, and IMS workloads.



Part	3090-600J six-way	9021-900 six-way	9021-9X2 ten-way	9672-RX3 ten-way	9672-RX4 ten-way	9672-RX5 ten-way
TCM/MCM	79	46	56	2	1	1
Processor boards	10	10	14	1	1	1
Processor chips	7300	4504	5960	86	35	31
Storage cards	160	48	48	4	4	4
Power supplies	157	154	202	12	12	15
Cables	1753	1153	1388	32	34	36
AMDs	42	78	87	4	6	6
Power/thermal/control cards	114	38	38	2	3	3

Figure 2

S/390 G3 and G4 servers vs. 10-way bipolar: (a) bipolar; (b) G3 server; (c) G4 server.

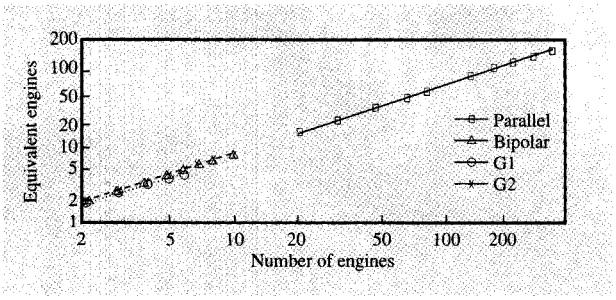
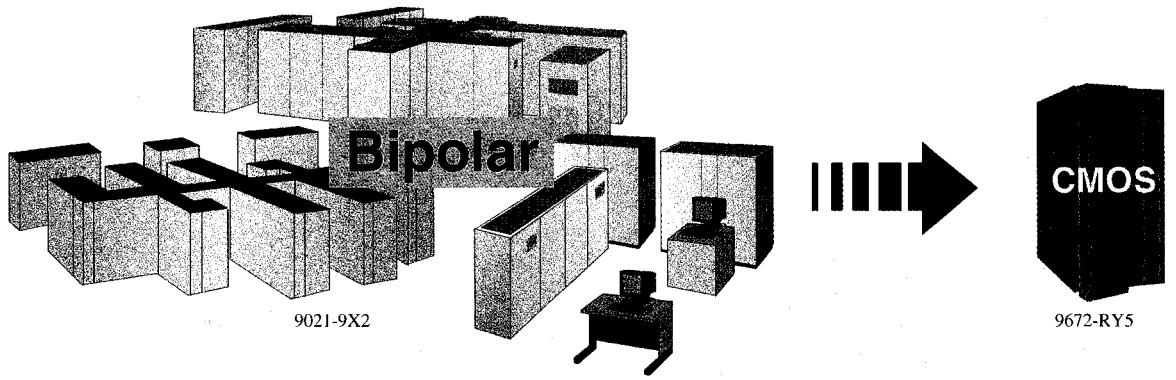
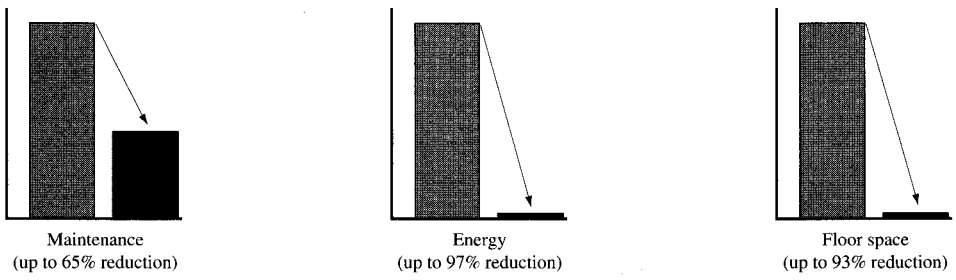


Figure 3

S/390 scalability: Parallel Sysplex vs. single server. (Parallel server: IMS environment; single server: Mixed LSPR workload.)

improve the total system availability for scheduled hardware maintenance (concurrent repair, such as $N + 1$ power supply channel cards, concurrent microcode patch apply, etc.) Along with the exploitation of parallel sysplex technology, S/390 provides the most nearly continuously available computing environment in the industry today. This is true especially from both hardware and software standpoints, especially the latter. In a parallel sysplex, not only can systems be switched in and out of the complex for hardware problems or upgrades without disruption; the same is true for software problems and maintenance. The customer can have near-continuous availability by using the IBM S/390 CMOS servers and parallel sysplex technology.





Measurement	9021-9X2	9672-RY5	Change (%)
Floor space (w/clearance) (ft^2)	671.6	51.9	93
Power (kVA)	144	5.0	97
Weight (lb)	31857	2067	93

Figure 4

Lower environmental costs: bipolar vs. CMOS.

Evolution of mainframe servers and network systems

• *Resurgence of mainframe servers*

IBM mainframes were physically large in comparison with pre-1990 systems, expensive to buy and maintain, difficult to operate, and limited to a host-centric domain. Without alternatives, large-system computer users were bound to that environment. However, despite these attributes, mainframes were still the marketplace's premiere platform for mission-critical applications. As long as such applications were profitable and provided a business advantage, IBM mainframes held a positive position in the marketplace.

By 1993, however, alternatives to the mainframe were clear, and the marketplace was ready to move. Client/server technologies were appearing that were smaller, less expensive, and easier to maintain and operate, with lower-cost, more productive, user-friendly, "total solutions" application software. An impending mainframe extinction was predicted. Reacting to this changing environment, IBM developed a mainframe transformation strategy which is currently successful. As a result, the resurgence of the mainframe is proceeding, although in a form that differs dramatically from the mainframe of old. Today's mainframe is smaller (Figure 2), less expensive to buy, less expensive to maintain (**Figure 4**) [3], and definitively more open. The reduction in physical size is a result of our transformation from bipolar to CMOS hardware and the associated change from water-cooled to air-cooled technology. The increased density of CMOS allows many more circuits per chip at a much lower power level, which also requires less cooling. The creation of OS/390 has reduced the cost of software and provided for openness that facilitates more competitive, leading-edge "total solution" applications [4]. Also, support provided for industry-standard connection devices such as Ethernet, ATM, and FDDI has enhanced the attractiveness of the mainframe. The parallel sysplex approach is the other key attribute of this successful transformation, since parallel sysplex allows for the clustering of far more processors in an integrated system than was practical in previous SSI multiprocessor systems.

Thus, the water-cooled bipolar mainframe of yesterday has been transformed to an air-cooled CMOS large-system server that is well positioned to meet the changing needs of the marketplace.

• *Requirement for microprocessor growth with industry*

The effect of combining competitive price/performance and open interfaces has been to attract many new applications to S/390 that were not originally written for it. Since many such applications are ported from UNIX/NT platforms and are generally not optimized for parallel sysplex, it is a key requirement to be able to scale their performance. This requires the SMP including the microprocessors to increase 40–50% [5] in performance every year.

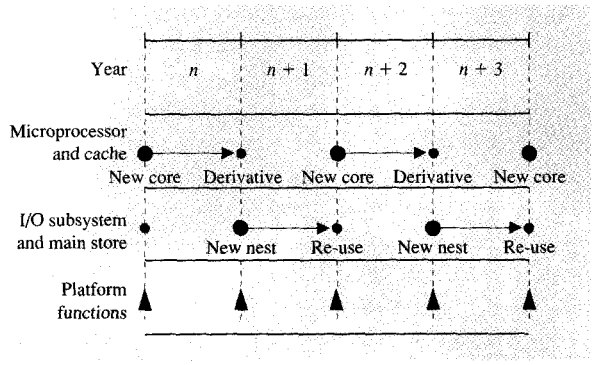


Figure 5

Development pipeline.

• *SMP design targets—staggered development pipeline*

Since the design of a new microprocessor in the industry takes about four years, it is necessary to structure the system differently from conventional designs in order to deliver 40–50% growth each year. The approach IBM has taken is to employ two microprocessor design teams that produce alternate microprocessor "cores" every two years, each with a technology upgrade in the interim years (**Figure 5**). The I/O and memory subsystems for the system are refreshed every two years, but staggered relative to the introduction of new microprocessor cores; this avoids the introduction of a new microprocessor with a new SMP environment, thereby simplifying the verification and architectural correctness of the resulting system.

• *I/O evolution*

To support the increasing processor performance of the S/390 CMOS servers and to provide more open, industry-standard interfaces, the S/390 I/O subsystem had to increase the number of parallel, ESCON*, and inter-system I/O channels attached to the server, and also had to develop a strategy to provide direct attachment of the new I/O interfaces. Increasing the total I/O bandwidth and connectivity to channels and new I/O adapters required the introduction of a new internal link called the self-timed interface (STI). The STI was introduced on G3 servers and provides high bandwidth at distances of the order of several meters. Since the STI requires relatively few chip I/O pins, many STIs are provided, further increasing the bandwidth and improving

the connectivity. While the STI provided the required connectivity and bandwidth, the attachment of more channels also required their redesign and remapping into state-of-the-art chip technologies. The newer chip technologies reduced the cost and physical size of the channels while improving their reliability and performance. Traditional S/390 I/O used external, channel-attached devices to provide open, industry-standard interfaces. Now, S/390 servers (Multiprise* 2000) have integrated adapters to provide Ethernet, token-ring, FDDI, and ATM interfaces. The ATM adapter is the first to be attached using internal peripheral component interconnect (PCI), and it demonstrates how new adapters may be attached in the future. Disk storage was also traditionally attached using external channel-attached devices, and the S/390 servers have developed adapters to integrate Small Computer System Interface (SCSI) attached disk drives. All of the changes described above have been made while retaining the S/390 I/O programming model. Future directions may augment this programming model to provide more direct I/O programming interfaces.

Summary

The transformation of the IBM S/390 line of mainframe computers has successfully reached a milestone: equivalence of CMOS-based and bipolar-based microprocessors. The current success of G3 and G4 is also due to the redefinition of the mainframe. The inclusion of key customer requirements such as scalability and support of open interfaces has made the new G3 and G4 offerings more competitive in the marketplace. Parallel Sysplex and OS/390 offerings have set the stage for the continued leadership and growth of S/390 products. This issue of the *IBM Journal of Research and Development* provides insight into some of the hardware development efforts that were an integral part of the transformation.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of X/Open Co., Ltd.

References

1. J. M. Nick, B. B. Moore, J.-Y. Chung, and N. S. Bowen, "S/390 Cluster Technology: Parallel Sysplex," *IBM Syst. J.* **36,** No. 2, 172–201 (1997).
2. G. M. King, D. M. Dias, and P. S. Yu, "Cluster Architectures and S/390 Parallel Sysplex Scalability," *IBM Syst. J.* **36,** No. 2, 221–241 (1997).
3. *A Business Strategy Report to Customers*, Order No. GF22-5005-00; 1997, available through IBM branch offices.
4. *S/390 Technology Leadership*, Order No. GF22-5008-00; 1997, available through IBM branch offices.
5. L. Gwennap, "Processor Performance Climbs Steadily," *Microprocessor Report* **9,** No. 1, January 23, 1995.

Received May 20, 1997; accepted for publication July 24, 1997

Gururaj S. Rao *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (gururao@vnet.ibm.com).*

Dr. Rao received his Bachelor of Engineering degree from the University of Mysore, India, his Master of Engineering degree from the Indian Institute of Science, India, and the Ph.D. degree from Stanford University, California, all in electrical engineering. He was an Assistant Professor of Electrical Engineering at Rice University, Houston, Texas, from 1975 to 1978. He joined the IBM Thomas J. Watson Research Center at Yorktown Heights, New York, in 1978, and worked on large-system processor structure studies. In 1983, Dr. Rao joined the Data Systems Division (now the System/390 Division) in Poughkeepsie, where he is currently the manager of the Processor Architecture and System Structure Department, with responsibility for developing future large-system requirements and architecture direction. Dr. Rao has received several academic honors as well as IBM awards. In 1991, he was appointed a Senior Technical Staff Member.

Thomas A. Gregg *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (GREGG at PKEDVM9, tomgregg@us.ibm.com).*

Mr. Gregg is a Senior Technical Staff Member in the S/390 System Design group. He received an SC.B. degree in engineering from Brown University in 1972 and continued his studies under a university fellowship, receiving an SC.M. degree in electrical engineering in 1974. He joined IBM at the Poughkeepsie Laboratory in 1973. Mr. Gregg has held various technical positions in the area of I/O subsystem design. He holds numerous patents utilized in IBM ESCON and Intersystem Channel products, and has received eight IBM Invention Achievement Awards. He received an IBM Outstanding Innovation Award and an IBM Corporate Award for work on ESCON products, and an IBM Outstanding Innovation Award for work on Intersystem Channel products.

Cyril A. Price *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (PRICECA at IBMUSM10, priceca@us.ibm.com).*

Dr. Price is Program Manager of CEC Subsystems in the S/390 Hardware Development Product Line Management organization, currently working on the development of the IBM S/390 CMOS servers. He graduated from Pratt Institute with a B.E.E. degree in 1969, and received his M.S.E.E. and D.E.E. degrees from Syracuse University in 1977 and 1986, respectively. He joined IBM in 1969 as a circuit/chip designer and then served two years as a U.S. Army Signal Corps officer. In 1978 he was named manager of Circuit and Subsystem Design. From 1979 to 1981, he was Manager of LSI Design and Manager of Exploratory Circuits. From 1981 to 1992, he held several management positions, including Advanced Design Manager responsible for CMOS chip development, and IBM Kingston Technology Manager responsible for technology support of IBM Enterprise System/9000 Type 9121 air-cooled processors. Dr. Price has received an IBM Outstanding Technical Achievement Award and an IBM Invention Achievement Award for patents and technical disclosures. In 1992 he was named Manager of Advanced Development, responsible for the initial research and development and successful product introduction of the IBM S/390 G4 microprocessor.

Chitta L. Rao *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (clrao@vnet.ibm.com).* Dr. Rao is with the S/390 platform management area, working on Parallel Sysplex performance. He received an M.S. in electrical engineering from McGill University, Montreal, Canada, and a Ph.D. in theoretical nuclear physics from the University of Tennessee. Prior to joining IBM, Dr. Rao was engaged in research in nuclear physics, digital filters, and image processing. He joined IBM to work on large-system scientific and engineering processor design and development. Since 1991 he has been involved in the performance analysis of coupling facility design alternatives and Parallel Sysplex system performance.

Steven J. Repka *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (srepka@vnet.ibm.com).* Mr. Repka is a Senior Engineer working in the S/390 System Design group. He received a B.S. in computer engineering from Case Western Reserve University in 1978, joining IBM at the Poughkeepsie Laboratory that same year. He has held a variety of technical and managerial positions in processor design, engineering systems test, and field product support. Mr. Repka has worked on the development of S/390 large systems, including the 3090 and ES/9000 processor families, and, most recently, the S/390 G4 server.